LEXICAL PHONOLOGY AND THE HISTORY OF ENGLISH

APRIL McMAHON

Department of Linguistics
University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011–4211, USA www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 2000

Printed in the United Kingdom at the University Press, Cambridge

Typeset in Times 10/13pt. [CE]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data
McMahon, April M. S.
Lexical Phonology and the history of English / April McMahon.
p. cm. – (Cambridge studies in linguistics; 91)
Includes bibliographical references and index.
ISBN 0 521 47280 6 hardback
1. English language – Phonology, Historical. 2. English language –
History. 3. Lexical Phonology. I. Title. II. Series.
PE1133.M37 2000
421′.5 – dc21 99–28845 CIP

ISBN 0 521 47280 6 hardback

Contents

Acknowledgements *page* xi

1 **The rôle of history** 1
1.1 Internal and external evidence 1
1.2 Lexical Phonology and its predecessor 5
1.3 Alternative models 13
1.4 The structure of the book 33

2 **Constraining the model: current controversies in Lexical Phonology** 35
2.1 Lexical Phonology and Morphology: an overview 35
2.2 Why constraints? Halle and Mohanan (1985) 50
2.3 Current controversies 53

3 **Applying the constraints: the Modern English Vowel Shift Rule** 86
3.1 Introduction 86
3.2 The Vowel Shift Rule and the Derived Environment Condition 88
3.3 Problems for lax-vowel Vowel Shift Rule 94
3.4 Problems for Level 1 Vowel Shift Rule 127

4 **Synchrony, diachrony and Lexical Phonology: the Scottish Vowel Length Rule** 140
4.1 Introduction 140
4.2 A brief external history of Scots and Scottish Standard English 141
4.3 The Scots dialects and Scottish Standard English: synchronic linguistic characteristics 145

4.4 Internal history 151
4.5 The Scottish Vowel Length Rule in Present-Day Scots and Scottish Standard English 170
4.6 From sound change to phonological rule 195

5 Dialect differentiation in Lexical Phonology: the unwelcome effects of underspecification 205
5.1 Introduction 205
5.2 Length, tenseness and English vowel systems 206
5.3 For and against the identity hypothesis 209
5.4 Underspecification 215

6 English /r/ 230
6.1 Introduction 230
6.2 English /r/: a brief outline 231
6.3 Non-rhotic /r/: an insertion analysis 234
6.4 Alternative analyses 247
6.5 Synchronic arbitrariness and diachronic transparency 264
6.6 Lexical Phonology and English /r/ 277
6.7 Retrospect and p[r]ospect 283

Bibliography 286
Index 302

1 *The rôle of history*

1.1 Internal and external evidence

Any linguist asked to provide candidate items for inclusion in a list of the slipperiest and most variably definable twentieth-century linguistic terms, would probably be able to supply several without much prompting. Often the lists would overlap (*simplicity* and *naturalness* would be reasonable prospects), but we would each have our own idiosyncratic selection. My own nominees are *internal* and *external evidence*.

In twentieth-century linguistics, types of data and of argument have moved around from one of these categories to the other relatively freely: but we can identify a general tendency for more and more types of evidence to be labelled *external*, a label to be translated 'subordinate to internal evidence' or, in many cases, 'safe to ignore'. Thus, Labov (1978) quotes Kuryłowicz as arguing that historical linguistics should concern itself only with the linguistic system before and after a change, paying no attention to such peripheral concerns as dialect geography, phonetics, sociolinguistics, and psycholinguistics. Furthermore, in much Standard Generative Phonology, historical evidence finds itself externalised (along with 'performance factors' such as speech errors and dialect variation), making distribution and alternation, frequently determined by introspection, the sole constituents of internal evidence, and thus virtually the sole object of enquiry. In sum, 'If we study the various restrictions imposed on linguistics since Saussure, we see more and more data being excluded in a passionate concern for what linguistics is *not*' (Labov 1978: 275–6).

Labov accepts that 'recent linguistics has been dominated by the drive for an autonomous discipline based on purely internal argument', but does not consider this a particularly fruitful development, arguing that 'the most notorious mysteries of linguistic change remain untouched by such abstract operations and become even more obscure' (1978: 277). He consequently pleads for a rapprochement of synchronic and diachronic study, showing that advances in phonetics and sociolinguistics, which have illuminated many aspects of change in progress, can equally explain completed changes, provided that we accept the uniformitarian principle: ‘that is, the forces which operated to produce the historical record are the same as those which can be seen operating today’ (Labov 1978: 281). An alliance of phonetics, sociolinguistics, dialectology and formal model-building with historical linguistics is, in Labov’s view, the most promising way towards understanding the linguistic past. We must first understand the present as fully as possible: ‘only when we are thoroughly at home in that everyday world, can we expect to be at home in the past’ (1978: 308).

Labov is not, of course, alone in his conviction that the present can inform us about the past. His own approach can be traced to Weinreich, Labov and Herzog’s (1968: 100) emphasis on ‘orderly heterogeneity’ in language, a reaction to over-idealisation of the synchronic system and the exclusion of crucial variation data. However, integration of the synchronic and diachronic approaches was also a desideratum of Prague School linguistics, as expressed notably by Vachek (1966, 1976, 1983). Vachek uses the term ‘external evidence’ (1972) to refer solely to the rôle of language contact and sociocultural factors in language change; this work has informed and influenced both contact linguistics and Labovian sociolinguistics. Although Vachek accepts external causation of certain changes, however, he still regards the strongest explanations as internal, involving the language’s own structure. This leads to attempts to limit external explanation, often via circular and ultimately unfalsifiable statements like Vachek’s contention (1972: 222) that ‘a language system … does *not* submit to such external influence as would be incompatible with its structural needs and wants’. For a critique of the internal/external dichotomy in this context, see Dorian (1993), and Farrar (1996).

More relevant to our discussion here is Vachek’s argument that synchrony is never truly static: ‘any language system has, besides its solid central core, its periphery, which need not be in complete accordance with the laws and tendencies governing its central core’ (1966: 27). Peripheral elements are those entering or leaving the system, and it is vital that they should be identified, as they can illuminate trends and changes in the system which would not otherwise be explicable, or even observable. Peripheral phonemes, for instance, might be those perceived as foreign; or have a low functional yield; or be distributionally restricted, like English /h/ or /ŋ/ (Vachek 1976: 178). A dynamic approach is therefore essential: the synchronically peripheral status of certain elements allows us to understand and perhaps predict diachronic developments, while the changes which have produced this peripherality can in turn explain irregularities in the synchronic pattern. This is not to say that Vachek collapses the two; on the contrary, his review of Chomsky and Halle (1968) is particularly critical of 'the lack of a clear dividing line that should be drawn between synchrony and diachrony' (1976: 307). Vachek considers Chomsky and Halle's extension of the Vowel Shift Rule from peripheral, learned forms like *serene* ~ *serenity*, to non-alternating, core forms like *meal*, an unjustified confusion of synchrony and diachrony: by in effect equating sound changes and synchronic phonological rules, Standard Generative Phonology in practice significantly reduces the useful conclusions which can be drawn about either.

Although Vachek seems to regard synchronic and diachronic data and analysis as mutually informing, the relationship is seen rather differently in Bailey's time-based or developmental linguistics. Bailey (1982: 154) agrees that 'any step towards getting rid of the compartmentalization of linguistics into disparate and incompatible synchronic, diachronic, and comparative or dialectal pursuits must ... be welcomed', and proposes polylectal systems sensitive to diachronic data. He coins the term 'yroëth' (which is theory spelled backwards) for 'something claiming to be a *theory* which may have a notation and terminology but fails to achieve any deep-level explanation ... All synchronic–idiolectal analysis is yroëthian, since deep explanation and prediction are possible only by investigating and understanding how structures and other phenomena have developed into what they have become' (Bailey 1996: 378). It is therefore scarcely surprising that Bailey regards the influence of diachronic on synchronic analysis as one-way, arguing that historical linguists are fundamentally misguided in adopting synchronic frameworks and notions for diachronic work: in doing so, they are guilty of analysing out the variation and dynamism central to language change by following the 'nausea principle': 'if movement makes the mandarins seasick, tie up the ship and pretend it is part of the pier and is not meant to sail anywhere' (Bailey 1982: 152).

We therefore have four twentieth-century viewpoints. The standard line of argumentation focuses on synchrony; historical evidence here is external, and is usable only as in Chomsky and Halle (1968), where sound changes appear minimally recast as synchronic phonological rules.

Vachek, conversely, argues that synchronic and diachronic phonology are equally valid and equally necessary for explanation. Labov argues that the present can tell us about the past, and Bailey the reverse. My own view is closest to Vachek's: if we are really to integrate synchrony and diachrony, the connection should cut both ways. That is, the linguistic past should be able to help us understand and model the linguistic present: since historical changes have repercussions on systems, an analysis of a synchronic system might sometimes benefit from a knowledge of its development. Perplexing synchronic phenomena might even become transparent in the light of history. But in addition, a framework originally intended for synchronic analysis will be more credible if it can provide enlightening accounts of sound change, and crucially model the transition from sound change to phonological rule without simply collapsing the two categories.

This book is thus intended as a contribution to the debate on the types of evidence which are relevant in the formulation and testing of phonological models, and has as one of its aims the discussion and eventual rehabilitation of external evidence. There will be particular emphasis on historical data and arguments; but issues of variation, which recent sociolinguistic work has confirmed as a prerequisite for many changes (Milroy and Milroy 1985; Milroy 1992), will also figure, and some attention will also be devoted to the phonetic motivation for sound changes and phonological rules.

However, although these arguments are of general relevance to phonologists, they are addressed here specifically from the perspective of one phonological model, namely Lexical Phonology. In short, the book also constitutes an attempt to constrain the theory of Lexical Phonology, and to demonstrate that the resulting model can provide an illuminating analysis of problematic aspects of the synchronic phonology of Modern English, as well as being consistent with external evidence from a number of areas, including diachronic developments and dialect differences. I shall focus on three areas of the phonology in which the unenviable legacy of Standard Generative Phonology, as enshrined in Chomsky and Halle (1968; henceforth SPE) seriously compromises the validity of its successor, Lexical Phonology: these are the synchronic problem of abstractness; the differentiation of dialects; and the relationship of sound changes and phonological rules. I shall show that a rigorous application of the principles and constraints inherent in Lexical Phonology permits an enlightening account of these areas, and a demonstration that generative models need not necessarily be subject to the failings and infelicities of their predecessor. Finally, just as the data discussed here are drawn from the synchronic and diachronic domains, so the constraints operative in Lexical Phonology will be shown to have both synchronic and diachronic dimensions and consequences.

1.2 Lexical Phonology and its predecessor

Lexical Phonology (LP) is a generative, derivational model: at its core lies a set of underlying representations of morphemes, which are converted to their surface forms by passing through a series of phonological rules. It follows that LP has inherited many of the assumptions and much of the machinery of Standard Generative Phonology (SGP; see Chomsky and Halle 1968). LP therefore does not form part of the current vogue for monostratal, declarative, non-derivational phonologies (see Durand and Katamba 1995, Roca (ed.) 1997a), nor is it strictly a result of the recent move towards non-linear phonological analyses, with their emphasis on representations rather than rules (see Goldsmith 1990, and the papers in Goldsmith (ed.) 1995). Although elements of metrical and autosegmental notation can readily be incorporated into LP (Giegerich 1986, Pulleyblank 1986), its innovations have not primarily been in the area of phonological representation, but rather in the organisational domain.

The main organisational claim of LP is that the phonological rules are split between two components. Some processes, which correspond broadly to SGP morphophonemic rules, operate within the lexicon, where they are interspersed with morphological rules. In its origins, and in the version assumed here, the theory is therefore crucially integrationist (but see Hargus and Kaisse (eds.) 1993 for discussion, and Halle and Vergnaud 1987 for an alternative view). The remainder apply in a postlexical, postsyntactic component incorporating allophonic and phrase-level operations. Lexical and postlexical rules display distinct clusters of properties, and are subject to different sets of constraints.

As a model attempting to integrate phonology and morphology, LP is informed by developments in both these areas. Its major morphological input stems from the introduction of the lexicalist hypothesis by Chomsky (1970), which initiated the re-establishment of morphology as a separate subdiscipline and a general expansion of the lexicon. On the phonological side, the primary input to LP is the abstractness controversy. Since the advent of generative phonology, a certain tension has existed between the desire for maximally elegant analyses capturing the greatest possible number of generalisations, and the often unfounded claims such analyses make concerning the relationships native speakers perceive among words of their language. The immensely powerful machinery of SGP, aiming only to produce the simplest overall phonology, created highly abstract analyses. Numerous attempts at constraining SGP were made (e.g. Kiparsky 1973), but these were never more than partially successful. Combating abstractness provided a second motivation for LP, and is also a major theme of this book.

The problem is that the SPE model aimed only to provide a maximally simple and general phonological description. If the capturing of as many generalisations as possible is seen as paramount, and if synchronic phonology is an autonomous discipline, then, the argument goes, internal, synchronic data should be accorded primacy in constructing synchronic derivations. And purely internal, synchronic data favour abstract analyses since these apparently capture more generalisations, for instance in the extension of rules like Vowel Shift in English from alternating to non-alternating forms. However, as Lass and Anderson (1975: 232) observe, 'it just might be the case that generalizations achieved by extraparadigmatic extension are specious'; free rides, for instance, 'may just be a property of the model, rather than of the reality that it purports to be a model of. If this should turn out to be so, then any "reward" given by the theory for the discovery of "optimal" grammars in this sense would be vacuous.' In contrast, I assume that if LP is a sound and explanatory theory, its predictions must consistently account for, and be supported by, external evidence, including diachronic data; the facts of related dialects; speech errors; and speaker judgements, either direct or as reflected in the results of psycholinguistic tests. This coheres with Churma's (1985: 106) view that '"external" ... data ... must be brought to bear on phonological issues, unless we are willing to adopt a "hocus pocus" approach ... to linguistic analyses, whereby the only real basis for choice among analyses is an essentially esthetic one' (and note here Anderson's (1992: 346) stricture that 'it is important not to let one's aesthetics interfere with the appreciation of fact'). The over-reliance of SGP on purely internal evidence reduces the scope for its validation, and detracts from its psychological reality, if we accept that 'linguistic theory ... is committed to accounting for evidence from all sources. The greater the range of the evidence types that a theory is capable of handling satisfactorily, the greater the likelihood of its being a "true" theory' (Mohanan 1986: 185).

These ideals are unlikely to be achieved until proponents of LP have the courage to reject tenets and mechanisms of SGP which are at odds with the anti-abstractness aims of lexicalism. For instance, although Mohanan (1982, 1986) is keen to stress the relevance of external evidence, he is forced to admit (1986: 185) that his own version of the theory is based almost uniquely on internal data. Elegance, maximal generality and economy are still considered, not as useful initial heuristics, but as paramount in determining the adequacy of phonological analyses (see Kiparsky 1982, Mohanan 1986, and especially Halle and Mohanan 1985). The tension between these relics of the SPE model and the constraints of LP is at its clearest in Halle and Mohanan (1985), the most detailed lexicalist formulation of English segmental phonology currently available. The Halle–Mohanan model, which will be the focus of much criticism in the chapters below, represents a return to the abstract underlying representations and complex derivations first advocated by Chomsky and Halle. Both the model itself, with its proliferation of lexical levels and random interspersal of cyclic and non-cyclic strata, and the analyses it produces, involving free rides, minor rules and the full apparatus of SPE phonology, are unconstrained.

Despite this setback, I do not believe that we need either reject derivational phonology outright, or accept that any rule-based phonology must inevitably suffer from the theoretical afflictions of SGP. We have a third choice; we can re-examine problems which proved insoluble in SGP, to see whether they may be more tractable in LP. However, the successful application of this strategy requires that we should not simply state the principles and constraints of LP, but must rigorously apply them. And we must be ready to accept the result as the legitimate output of such a constrained phonology, although it may look profoundly different from the phonological ideal bequeathed to us by the expectations of SGP.

In this book, then, I shall examine the performance of LP in three areas of phonological theory which were mishandled in SGP: abstractness; the differentiation of related dialects; and the relationship of synchronic phonological rules and diachronic sound changes. If LP, suitably revised and constrained, cannot cope with these areas adequately, it must be rejected. If, however, insightful solutions can be provided, LP will no longer be open to many of the criticisms levelled at

SGP, and will emerge as a partially validated phonological theory and a promising locus for further research.

The three issues are very clearly connected; let us begin with the most general, abstractness. SGP assumes centrally that the native speaker will construct the simplest possible grammar to account for the primary linguistic data he or she receives, and that the linguist's grammar should mirror the speaker's grammar. The generative evaluation measure for grammars therefore concentrates on relative simplicity, where simplicity subsumes notions of economy and generality. Thus, a phonological rule is more highly valued, and contributes less to the overall complexity of the grammar, if it operates in a large number of forms and is exceptionless.

This drive for simplicity and generality meant exceptions were rarely acknowledged in SGP; instead, they were removed from the scope of the relevant rule, either by altering their underlying representations, or by applying some 'lay-by' rule and a later readjustment process. Rules which might be well motivated in alternating forms were also extended to non-alternating words, which again have their underlying forms altered and are given a 'free ride' through the rule. By employing strategies like these, a rule like Trisyllabic Laxing in English could be made applicable not only to forms like *divinity* (~ *divine*) and *declarative* (~ *declare*), but also to *camera* and *enemy*; these would have initial tense vowels in their underlying representations, with Trisyllabic Laxing providing the required surface lax vowels. Likewise, an exceptional form like *nightingale* is not marked [−Trisyllabic Laxing], but is instead stored as /nɪxtVngǣl/; the voiceless velar fricative is later lost, with compensatory lengthening of the preceding vowel, to give the required tense vowel on the surface.

The problem is that the distance of underlying representations from surface forms in SGP is controlled only by the simplicity metric – which positively encourages abstractness. Furthermore, there is no linguistically significant reference point midway between the underlying and surface levels, due to the SGP rejection of the phonemic level. Consequently, as Kiparsky (1982: 34) says, SGP underlying representations 'will be at *least* as abstract as the classical phonemic level. But they will be more abstract whenever, and to whatever extent, the simplicity of the system requires it.' This potentially excessive distance of underliers from surface forms raises questions of learnability, since it is unclear how a child might acquire the appropriate underlying representation for a non-alternating form.

A further, and related, charge is that of historical recapitulation: Crothers (1971) accepts that maximally general rules reveal patterns in linguistic structure, but argues that these generalisations are non-synchronic. If we rely solely on internal evidence and on vague notions of simplicity and elegance to evaluate proposed descriptions, we are in effect performing internal reconstruction of the type used to infer an earlier, unattested stage of a language from synchronic data. Thus, Lightner (1971) relates *heart* to *cardiac* and *father* to *paternal* by reconstructing Grimm's Law (albeit perhaps not wholly seriously), while Chomsky and Halle's account of the *divine* ~ *divinity* and *serene* ~ *serenity* alternations involves the historical Great Vowel Shift (minimally altered and relabelled as the Vowel Shift Rule) and the dubious assertion that native speakers of Modern English internalise the Middle English vowel system. I am advocating that historical factors should be taken into account in the construction and evaluation of phonological models; but the mere equation of historical sound changes and synchronic phonological rules is not the way to go about it.

Here we confront our second question: how are sound changes integrated into the synchronic grammar to become phonological rules? In historical SGP (Halle 1962, Postal 1968, King 1969), it is assumed that a sound change, once implemented, is inserted as a phonological rule at the end of the native speaker's rule system; it moves gradually higher in the grammar as subsequent changes become the final rule. This process of rule addition, or innovation, is the main mechanism for introducing the results of change into the synchronic grammar: although there are occasional cases of rule loss or rule inversion (Vennemann 1972), SGP is an essentially static model. The assumption is that underlying representations will generally remain the same across time, while a cross-section of the synchronic rule system will approximately match the history of the language: as Halle (1962: 66) says, 'the order of rules established by purely synchronic considerations – i.e., simplicity – will mirror properly the relative chronology of the rules'. Thus, a sound change and the synchronic rule it is converted to will tend to be identical (or at least very markedly similar), and the 'highest' rules in the grammar will usually correspond to the oldest changes. SGP certainly provides no means of incorporating recent discoveries on sound change in progress, such as the division of diffusing from non-diffusing changes (Labov 1981).

It is true that some limited provision is made in SGP for the restructuring of underlying representations, since it is assumed that children will learn the optimal, or simplest, grammar. This may not be identical to the grammar of the previous generation: whereas adults may only add rules, the child may construct a simpler grammar without this rule but with its effects encoded in the underlying representations. However, this facility for restructuring is generally not fully exploited, and the effect on the underliers is in any case felt to be minimal; thus, Chomsky and Halle (1968: 49) can confidently state:

> It is a widely confirmed empirical fact that underlying representations are fairly resistant to historical change, which tends, by and large, to involve late phonetic rules. If this is true, then the same system of representation for underlying forms will be found over long stretches of space and time.

This evidence that underlying representations are seen in SGP as diachronically and diatopically static, is highly relevant to our third problem, the differentiation of dialects. The classical SGP approach to dialect relationships therefore rests on an assumption of identity: dialects of one language share the same underlying representations, with the differences resting in the form, ordering and/or inventory of their phonological rules (King 1969, Newton 1972). Different languages will additionally differ with respect to their underlying representations. The main controversy in generative dialectology relates to whether one of the dialects should supply underlying representations for the language as a whole, or whether these representations are intermediate or neutral between the realisations of the dialects. Thomas (1967: 190), in a study of Welsh, claims that 'basal forms are *dialectologically mixed*: their total set is not uniquely associated with any total set of occurring dialect forms'. Brown (1972), however, claims that considerations of simplicity compel her to derive southern dialect forms of Lumasaaba from northern ones.

This requirement of a common set of underlying forms is extremely problematic (see chapter 5 below). Perhaps most importantly, the definition of related dialects as sharing the same underlying forms, but of different languages as differing at this level, prevents us from seeing dialect and language variation as the continuum which sociolinguistic investigation has shown it to be. Furthermore, the family tree model of historical linguistics is based on the premise that dialects may diverge across time and become distinct languages, but this pattern is obscured by the contention that related dialects are not permitted to differ at the underlying level, while related languages characteristically do. It is not at all clear what conditions might sanction the sudden leap from a situation where two varieties share the same underlying forms and differ in their rule systems, to a revised state involving differences at all levels. These theoretical objections are easily swept aside, however, in a model like SGP where the central assumptions require maximal identity in the underlying representations.

The three areas outlined above are all dealt with unsatisfactorily in SGP; moreover, these deficiencies are due in all cases, directly or indirectly, to the insistence of proponents of the SPE model on a maximally simple, exceptionless phonology. The use of an evaluation measure based on simplicity, the lack of a level of representation corresponding to the classical phonemic level, and the dearth of constraints on the distance of underlying from surface representations all encourage abstractness. Changes in the rule system are generally preferred, in such a system, to changes in the underlying forms, which are dialectally and diachronically static. Rules simply build up as sound changes take effect, with no clear way of encoding profound, representational consequences of change, no means of determining when the underliers should be altered, and no link between sound changes and phonological rules save their identity of formulation. This historical recapitulation contributes to further abstractness, and means that, in effect, related dialects *must* share common underlying forms. King (1969: 102) explicitly states that external evidence, whether historical or from related dialects, may play no part in the evaluation of synchronic grammars; this is presented as a principled exclusion, since speakers have no access to the history of their language or to the facts of related varieties, but is equally likely to be based on the clear inadequacies of SGP when faced with data beyond the synchronic, internal domain.

I hope to show in the following chapters that LP need not share these deficiencies, and that its successes in the above areas are also linked. Working with different varieties of Modern English, I shall demonstrate that the abstractness of the synchronic phonology can be significantly restricted in LP. In general, the strategy to be pursued will involve imposing and strengthening the constraints already existing in LP, most notably the Strict Cyclicity Condition or Derived Environment Condition, and assessing the analyses which are possible, impossible, or required within the constrained model. Because maximally surface-true analyses will be enforced for each variety, we will be unable to consistently derive related dialects from the same underlying representations, and the underliers will also be subject to change across time. Sound changes and related phonological rules will frequently differ in their formulation, and new links between diachrony and synchrony will be revealed.

Of course, this is not the first time that questions have been raised over aspects of SGP: for instance, I have already quoted Lass and Anderson (1975), a Standard Generative analysis of Old English phonology incorporating an extremely eloquent and perceptive account of the difficulties which seemed then to face SGP, a model which had seemed so 'stable and unified' (1975: xiii) in 1970, when their account of Old English was first drafted. Lass and Anderson set out to test SGP against a particular set of data. They discover that the theory makes particular predictions; that it permits, or even requires, them to adopt particular solutions. These solutions are sometimes fraught with problems. Lass and Anderson could, of course, have made use of the power of SGP to reformulate the areas where they identify problems and weaknesses; instead, they include a final section explicitly raising doubts about the theory, and the issues they identify have been crucial in remodelling phonological theory ever since.

The conclusion, more than twenty years on, is that these difficulties cannot be solved within SGP: the simplicity metric, the overt preference (without neurological support) for derivation over storage, and the denial of 'external' evidence, mean that many of the generalisations captured are simply over-generalisations. The model must be rejected or very radically revised.

LP is one result. But the revisions have so far not been radical enough. I shall show in the following chapters that it is possible to maintain the core of the generative enterprise in phonology (namely, that alternating surface forms may be synchronically derived from a common underlier) without a great deal of the paraphernalia which was once thought to be crucial to the goal of capturing significant generalisations, but in practice encouraged the statement of artefactual and insignificant ones. Thus, we shall reject the SGP identity hypothesis on dialect variation; rule out free rides; prohibit derivation in non-alternating morphemes; revise the feature system; and exclude underspecification, which has recently become an expected ingredient of LP, but is in fact quite independent from it.

In the rest of the book, then, I shall follow much the same route as Lass and Anderson: we shall begin with a phonological model, in this case LP, and assess its performance given a particular set of data, here the vowel phonology, loosely defined, of certain accents of Modern English. The model is characterised by a number of constraints; I shall argue that these should be rigorously applied, and indeed supplemented with certain further restrictions. We can then examine what is possible within the model, and what solutions it forces us to adopt. If we are forced to propose analyses which seem to conflict with internal or external evidence, being perhaps apparently unlearnable, or counter-historical, or without phonetic or diachronic motivation, we must conclude that the model is inadequate. Likewise, the model may never make decisions for us: in other words, any analysis may be possible. Such a theory clearly makes no predictions, and is unconstrained, unfalsifiable and uninteresting. On the other hand, we may find that the predictions made are supported by internal and external evidence; that the phonology becomes more concrete, and arguably more learnable than the standard model; that phonetics and phonology can be better integrated, and the relationship between them better understood; and that a more realistic model of variation and change can be proposed.

So far, I have introduced LP only in the broadest terms. A number of outlines of LP are available (Kiparsky 1982, 1985; Mohanan 1982, 1986; Pulleyblank 1986; Halle and Mohanan 1985). However, most aspects of LP, including its central tenets, are still under discussion (see Hargus and Kaisse (eds.) 1993, Wiese (ed.) 1994). Available introductions therefore tend to be restricted to presenting the version of LP used in the paper concerned (Kaisse and Shaw 1985 does provide a broader perspective, but is now, in several crucial respects, out of date). Consequently, it may be difficult for a reader not entirely immersed in the theory to acquire a clear idea of the current controversies, which become apparent only by reading outlines of LP incorporating opposing viewpoints. I shall consequently attempt in chapter 2 to provide an overview of LP, considering both its evolution, and current controversies within the theory which will be returned to in subsequent chapters. First, however, I must justify approaching the problems outlined above in a derivational model at all.

1.3 Alternative models

Sceptical observers, and non-generative phonologists, may see my programme as excessively idealistic, on the not unreasonable grounds that generative phonology is by its very nature far too flexible to allow adequate constraint. In other words, given phonological rules and underlying forms, an analysis can always be cobbled together which will get the right surface forms out of the proposed underliers: if the first attempt doesn't do the trick, you can alter the underliers, or the rules, until you find a set-up that works. And since LP is generative, and phonologists are no less ingenious now than in the heyday of SGP, the new model is open to precisely the same criticism as the old one. Here again, Lass and Anderson (1975: 226) ask: 'But is the mere fact that a phonological solution works any guarantee that it is correct?' Of course not: it is precisely because we cannot rely purely on distribution and alternation that we need extra, 'external' evidence. The analyses I shall propose in subsequent chapters will look peculiar in SGP terms; but I hope to show that they are coherent with evidence of a number of different kinds, and that they allow interesting predictions to be made. For instance, we shall see that my analysis of the English Vowel Shift specifies a principled cut-off point between what can be derived, and what cannot, giving a partial solution to the determinacy problem. A typical progression from sound changes to phonological rules will also be identified, giving a certain amount of insight into variation and change, as well as the embedding of change in the native speaker's grammar. These implications and conclusions lend support to LP, and suggest, if nothing else, that the model should be pursued and tested further. Phonetics, phonology, variation and change cannot be integrated in this way in SGP. I have not yet seen similar clusterings of evidence types in non-generative phonologies, either.

Arguments of this kind give me one reason for adopting LP, and attempting to constrain generative phonology, rather than rejecting a derivational model altogether. Nonetheless, questions will undoubtedly be raised concerning the relevance of this work, given the current move towards monostratal, declarative, and constraint-based phonologies. I cannot fully address these issues here, but the rest of the book is intended as a partial answer; and I also have some questions of my own.

1.3.1 Rules and constraints

Let us begin with the issue of rules versus constraints (see Goldsmith (ed.) 1993a, and Roca (ed.) 1997a). There seems to be a prevailing opinion in current phonology that it is somehow more respectable to work with constraints only, than to propose rules and then constrain their application, however heavily. For instance, Government Phonology

(Kaye, Lowenstamm and Vergnaud 1985, Kaye 1988) includes principles and parameters, but no destructive operations, while Optimality Theory (Prince and Smolensky 1993) incorporates only constraints.

We might assume that positing constraints *per se* is uncontroversial, as they are part of all the phonological models surveyed here: but they are still criticised when they are part of theories which also contain rules, like LP. For instance, Carr (1993: 190–1) accepts that LP may in principle be highly constrained and therefore relatively non-abstract, but argues that 'The crucial issue here is whether such constraints (if they are desirable) come from within the theory or have to be imposed from outside. If the latter is the case, then the LP theory itself is, for those seeking a non-abstract phonology, in need of revision.' How are we to assess whether constraints are 'imposed from outside'? Is the condition against destructive operations in Government Phonology not 'imposed from outside'? Why should the specification of the number of vowel or consonant elements, or the assumption that reference should be made to universal, innate principles, have the status of internally determined, intrinsic aspects of the theory, while the constraints of LP should not? For example, I shall argue below that the main constraint on LP is the Strict Cyclicity Condition (SCC), which does follow from the architecture of the model, insofar as it is restricted to the (universally cyclic) first lexical level. Moreover, it is quite possibly derivable from the arguably innate Elsewhere Condition, and may not therefore require to be independently stated. Even so, why should this be seen as such a conclusive advantage? If we consider language change, we see that purely formal attempts to explain developments have rarely been very successful. For instance, in the domain of word order change, scholars like Lehmann (1973) and Vennemann (1974) attempted to account for the correlations of certain logically independent word order properties, and the fact that the change of one often seemed to have repercussions for others, in terms of the principle of natural serialisation; this would probably be interpreted today as a principle or a parameter (see Smith 1989). However, this principle is not, on its own, explanatory (Matthews 1981): it is only when issues of parsing and learnability (see Kuno 1974) are invoked that we begin to understand why change should proceed so regularly in a particular direction. It seems highly likely that the same should be true of phonology: synchronically or diachronically, we need external evidence to explain why certain patterns occur and recur. Thus, the SCC is not purely a formal constraint. Instead, like Kiparsky's

Alternation Condition (Kiparsky 1973), which it is partially intended to formalise, it is a learnability constraint: grammars violating either condition will be harder to learn. This means that, for instance, a grammar ordering rules on Level 1, within the domain of SCC, should be easier to acquire than a similar grammar with the same rules permitted to apply on Level 2, where they will not be controlled by SCC.

However, there is one crucial difference between the constraints of LP and those of Optimality Theory, for instance: the former restrict rule applications, whereas the latter replace rules. The next question, then, is whether rules are required at all. There are two considerations here, which relate in turn to the question of transparency in the synchronic grammar, and to the importance accorded to universality.

Anderson (1981), in a study of 'Why phonology isn't "natural"', argues that the effects of sound changes may build up in a language over time so that ultimately extremely opaque phonological processes may be operating synchronically. For instance, in Icelandic, Velar Fronting operates in a synchronically highly peculiar environment, giving back velars before the front vowels [y] and [ø], and front velars before the diphthong [ai], with a back first element. However, once we know that historically, the problematic front vowels are from back [u] and [ɔ], while the difficult diphthong was earlier front [æ:], we can see that Velar Fronting applies in the context of historically front vowels. Anderson points out that a synchronic grammar must nonetheless contain a description of these facts, and that this synchronic rule will not be phonetically motivated, or universal. The synchronic state is simply the result of language-specific history, and the fact that we have a historical explanation means the synchronic rule need provide no more than a description.

Everyday, work-horse descriptive work of this language-particular kind is what phonological rules are for, and it is my contention that phonological theories need them, whether their proponents are happy to admit it or not. For instance, Goldsmith's introduction to his (1993) collection of papers, entitled *The Last Phonological Rule*, argues that rules and derivations should not be part of a theory of phonology. However, Hyman's (1993) paper, despite setting out to find cases where extrinsic rule ordering will not work, comes to the conclusion that it is, in fact, a viable approach, while other papers (notably Goldsmith's and Lakoff's) involve language-specific constraints, such as Lakoff's (1993: 121) statement that 'When C precedes ʔ# at level W, an /e/ absent at level

W intervenes at level P', which is surely an epenthesis rule by any other name. As Padgett (1995) notes, these papers also include sequential, extrinsic level-ordering of constraints, and are therefore scarcely free of the apparatus of derivational phonology.

Similarly, Coleman (1995: 344) argues that 'Far from being a rule free theory completely unlike the SPE model, as its proponents claim, Government-based phonological analyses employ various derivational devices which are transformational rules in all but name ... Government Phonology is therefore as unconstrained as the models it seeks to replace.' For instance, Coleman points out that, to model the ostensibly prohibited deletion of segments, Government Phonology can first delete each marked element in turn, which the theory will permit; this will ultimately leave only the single 'cold' element which can be removed by the Obligatory Contour Principle (OCP) (see also 1.3.2 below). Furthermore, many of the principles invoked in Government Phonology seem language-specific; for instance, as we shall see in chapter 6, Harris (1994) argues that the loss of [r] in non-rhotic English dialects results from the innovation of the Non-Rhoticity Condition, which allows the R element to be licensed only in onsets. This condition allows an accurate description of the synchronic situation: the question is why such a constraint should become operative in the grammar of a particular dialect or set of dialects at a particular time. We might be dealing with a parameter resetting; but then, of course, we would have to ask why the resetting happened. Principles and parameters theory is faced with similar difficulties in historical syntax; thus, Lightfoot (1991: 160) remarks that, at the point when a parameter is reset, 'an abrupt change takes place, but it was preceded by gradual changes affecting triggering experiences but not grammars'. So, Lightfoot recognises 'piecemeal, gradual and chaotic changes' in the linguistic environment; these can affect, for instance, the frequency of a construction, and may be introduced for reasons of contact, or for stylistic effect. These changes are not amenable to systematic explanation; but they are important in creating the conditions for parameter resetting, which *is* intended to be explicable in terms of Lightfoot's theory of grammar. It is quite unclear where the language change actually begins, and what the status of these preparatory changes is. Of course, a rule-based theory has no particular advantages here; a rule of [r]-deletion would simply be written as a response to the loss of a segment which was present before, and we would seek out reasons for the loss in, for instance, phonetics or sociolinguistics. But we would not be taking the portentous step of labelling this variety-specific behaviour as a condition or a constraint, or falsely implying universality.

Finally, and most controversially, we turn to Optimality Theory (OT). In this theory, Universal Grammar for phonology consists of two components, a function Gen, and a set of universal constraints on representational well-formedness. Gen (for 'generate') takes a particular input, which will be a lexical entry, and generates *all possible outputs* – an infinite set of possible candidate analyses, which is then evaluated by the list of constraints. These constraints are universal, but crucially ordered differently for each language, to give the different attested surface results. Most theories of constraints in phonology have held that constraints are exceptionless. In OT, every constraint is potentially violable. This means that the 'winning', or maximally harmonic representation will not necessarily be the one which satisfies every constraint. It will be the one which violates fewest. More accurately, since constraints are ranked, it will be the candidate parse which violates fewest high-ranking constraints.

Prince and Smolensky (1993: 101) accept that 'Any theory must allow latitude for incursions of the idiosyncratic into grammar.' However, they argue that idiosyncratic behaviour is not modellable using rules, but rather by '(slightly) modified versions of the universal conditions on phonological form out of which core grammar is constructed ... [which] interact with other constraints in the manner prescribed by the general theory' (ibid.). This assumption has various consequences. First, constraints may be too low-ranked in particular languages to have any discernible effect. This is not taken to affect learnability adversely, since the strong assumption of universality means the constraints do not have to be learned, only their ranking; note, however, that acquisition is non-trivial given the explosion of constraints to be ranked in recent versions of the theory: Sherrard (1997) points out that only five constraints will give 120 possible grammars, while ten will allow 36 million. Contrast this with a rule-based approach, where a rule is written only where it captures phonological behaviour in the language concerned; we would not write, for instance, a universal version of the Vowel Shift Rule with effects tangible only in English and concealed elsewhere. To do so would be against every requirement of learnability, and would also unacceptably blur the distinction between the universal and the language-specific.

However, the question also arises of quite how different a constraint-based theory like OT is from a rule-based one. Prince and Smolensky's contention that constraints can be language-specifically modified leads to formulations like the now notorious Lardil FREE-V (1993: 101), which states that 'word-final vowels must not be parsed (in the nominative)', and again seems a static recasting of a very language-specific deletion rule. In similar vein, Prince and Smolensky (1993: 43), in considering the constraint NONFINALITY, note that 'It remains to formulate a satisfactory version of NONFINALITY for Latin.' What this means is that, logically, the issue is not solely one of determining the place of constraint C in the hierarchy of Language X. The formulation of C may also differ, and it is not clear how appreciably, between Languages X and Y. More generally, there is an issue of extrinsic ordering here, since while many constraints must be ranked language-specifically, there are others which are never violated, and which must therefore be placed universally at the top of the hierarchy. Prince and Smolensky (1993: 46) argue that this is acceptable since 'we can expect to find principles of universal ranking that deal with whole *classes* of constraints'. If ordering is acceptable when it refers to classes of ordered items, a rule-based model should be equally highly valued provided that it involved level-ordering, or ordering all lexical before all postlexical rules, for instance.

Even closer to the core of OT, the definition of the function Gen is itself controversial. Although Prince and Smolensky (1993: 79) advocate a parallel interpretation, they concede that Gen can also be understood serially, in which case its operation is much closer to a conventional derivation:

> some general procedure (DO-α) is allowed to make a certain single modification to the input, producing the candidate set of all possible outcomes of such modification. This is then evaluated; and the process continues with the output so determined. In this serial version of grammar, the theory of rules is narrowly circumscribed, but it is inaccurate to think of it as trivial.

However, this serial interpretation of Gen may be necessary; Blevins (1997) argues strongly that, without it, there is no way of verifying constraint tableaux, as each tableau will contain the allegedly maximally harmonic parse plus a random set of other candidates, but will not contain all possible parses, and therefore crucially does not contain all the evidence necessary to permit evaluation.

The perceived advantage of an OT account is the absence of specific processes; but it is unclear why such a theory, with vast overgeneration courtesy of Gen, should be seen as more parsimonious than a derivational theory with a finite number of non-overgenerating language-specific rules. Of the papers in Roca (ed.) (1997a), which focus on the rules–constraints debate, a surprising number contend that rules and derivations are still necessary, while Roca himself notes that 'OT is stretching its original formal fabric in ways that closer scrutiny may reveal are nothing but covert rules, and perhaps even derivations' (1997b: 39). Indeed, some work in OT is entirely open about the addition of rules: McCarthy (1993: 190) includes an epenthesis rule to account for the distribution of English /r/, and states quite explicitly that 'By a "rule" here I mean a phonologically arbitrary stipulation: one that is outside the system of Optimality.' As Halle and Idsardi (1997: 337–8) argue, 'Conceptually, reliance on an arbitrary stipulation that is outside the system of Optimality is equivalent to giving up on the enterprise. Data that cannot be dealt with by OT without recourse to rules are fatal counter-examples to the OT research programme.' At the very least, this introduction of rules alongside constraints removes the alleged formal superiority of OT, making it just as theoretically heterogeneous as LP, for instance, in containing both categories of statement.

1.3.2 Modelling sound changes

We return now more specifically to diachronic evidence. Proponents of some recent phonological models explicitly exclude historical processes from their ambit; Coleman (1995: 363), for instance, working within Declarative Phonology, refuses to consider one of Bromberger and Halle's (1989) arguments for rule ordering because of 'its diachronic nature. The relevance of such arguments to synchronic phonology is highly controversial, and thus no basis on which to evaluate the transformational hypothesis.' I reject this curtailment of phonological theory for two reasons. First, more programmatically, theorists should not be able to decide *a priori* the data for which their models should and should not account. It is natural and inevitable that a model should be proposed initially on the basis of particular data and perhaps data types, but it is central to the work reported below that the model subsequently gains credence from its ability to deal with quite different (and perhaps unexpected) data, and loses credibility to the extent that it fails with respect to other evidence. Secondly, and more pragmatically, no absolute distinction can be made between synchronic and diachronic phonology. Variation is introduced by change, and in turn provides the input to further change; and even if we are describing a synchronic stage, we must